# BUILDING A CUSTOM HIGH THROUGHPUT PLATFORM AT THE JOINT GENOME INSTITUTE FOR DNA CONSTRUCT DESIGN AND ASSEMBLY – PRESENT AND FUTURE CHALLENGES

Ian K. Blaby and Jan-Fang Cheng

## Supplemental File

BIOINFORMATIC TOOLS OVERVIEW

Our bioinformatic software comprises BOOST, gRNA-SeqRET, BLiSS and SynTrack. Each tool has been specifically designed and built to support the workflow and throughput of the synthesis platform at JGI. The Build Optimization Software Tools (BOOST) pipeline consists of a string of algorithms enabling redesign of provided DNA sequence for synthesis (*1*). Since no technology currently exists that is capable of synthesizing any conceivable DNA sequence, it is often necessary to exploit redundancy in the genetic code by altering codon usage without affecting the encoded protein sequence to overcome synthesis constraints, such as repetitive sequence, extreme GC/AT-skew or palindromes. BOOST offers the ability to iteratively refactor the codon usage ("polishing") and test against either local rules or direct interaction with a vendor's server via an API until synthesizable sequences are accomplished. BOOST also implements multiple strategies for codon optimization, as well as the automated designs for scar-less re-joining of sequences too long to be synthesized as a single fragment and assembly into any given vector by Gibson assembly. BOOST is hosted at the DOE National Energy Research Scientific Computing Center (NERSC) and is freely available for non-commercial use at http://boost.jgi.doe.gov . Directions on usage are available under the "manual" tab, and a webinar tutorial is available at https://j5.jbei.org/Videos/BOOST_3_16_2020.mp4 .

gRNA-SeqRET builds added functionality to CCTop (*29*), enabling the extraction of sections of genomic DNA (with coding regions defined vs. non-coding regions) of prescribed length and coordinates (Simirenko et al., in preparation). This tool has multiple functionalities, for example it can be used to select whole or partial genome gRNA targets based on user input parameters for CRISPR-mediated functional screens, to extract all potential promoter sequences from an annotated genome, or to extract homology arms for gene knock-in and knock-out designs.

All requested synthesis DNA sequences are filtered by the Black List Sequence Screening (BLiSS) tool prior to synthesis (Simirenko et al., in preparation). This application establishes sequences of potential biosecurity risk by identifying regions of significant homology on the Select agents or toxins list, defined by the Department of Health and Human Services (HHS) and Centers for Disease Control and Prevention (CDC), and allows researchers to be forewarned of flagged sequences of possible concern.

Although several platforms exist for tracking and managing synthetic biology workflows (*2*), we are developing an in-house solution called SynTrack. SynTrack performs as both tracking software (LIMS) and a workflow manager (Supplemental Figure 1). Project-specific data are entered into the database at the point of initiation of a project enabling the progress of each entity to be tracked through to completion or to merge partially completed workflows. Additionally, SynTrack aids in the automation workflow of the platform, for example, by specifying coordinates for oligonucleotides in plates for amenability to liquid handling workflows.

## MOLECULAR WORKFLOWS

The platform endeavors to work with successful proposals to achieve the users scientific goals, with no specific limitations on project types so long as they fit broadly to the platform portfolio

(Figure 1). For example, while limiting constructs to one or a handful of standard vectors would increase our throughput, this restraint would impede research by necessitating sub-cloning in the user's laboratory. Equally, cloned insert sizes range from a few hundred bp to >60kb for gRNA libraries and biosynthetic gene clusters respectively.

Due to this versatility the platform employs a more open-ended approach, with no single workflow for a given product type (Figure 1; Supplemental Figure 2). There are however guidelines that are followed, as indicated below and in Supplemental Figure 2. Since our current primary vendor has a maximum non-clonal size of 1800bp, and assembly methods differ in efficiency depending on the number of fragments, the total insert size defines how constructs are assembled. Of note is that while 1800bp is offered, we restrict orders to a maximum 1700bp as we have found this significantly reduces the error rate. For details on how assemblies are designed, and inserts >1700bp are designed to be partitioned into multiple fragments, see the Bioinformatic tools overview section above.

Inserts totaling <7kb, equating to ~5 fragments plus vector backbone, are always assembled via Gibson assembly. By default, we use 30 bases of homology to mediate the chewback and assembly reactions for Gibson assembly, which are automatically defined by BOOST (*1*). All DNA fragments are PCR amplified to both increase starting material and remove DNA linkers added by the vendor as part of their chemical synthesis process; the design of the oligonucleotides required for this amplification is automated as part of the BOOST workflow. Our preferred DNA polymerase and Gibson assembly mix is Kapa HIFI HotStart and Gibson Assembly Mastermix (Kapa Biosystems and New England Biosciences respectively).

In our experience, the efficiency of >5 inserts by Gibson assembly reduces significantly with every additional fragment, which is not compatible with the production-scale of construct assembly (repeating failed assemblies significantly impacts platform throughput). Therefore, for larger

inserts comprising 5 or more fragments we typically employ yeast assembly (TAR (*3*)), where (in our hands) up to 10 inserts can be efficiently assembled. If the destination vector provided by the user does not already contain a yeast origin or replication and selection marker the plasmid must first be modified to incorporate an appropriate cassette. Assembly on this scale allows for total insert size approaching 20kb. However, for final inserts of this magnitude we purchase cloned-inserts rather than linear strings (which can be excised by restriction digest) for which the current maximum size is 5000bp from our primary vendor. The reason is that for this number of inserts the likelihood of a mutation being introduced by the polymerase reaction increases, whereas purchasing cloned inserts allows for restriction digest excision of the insert avoiding PCR reactions. For insert sizes exceeding 20kb, such as for biosynthetic gene clusters, we perform multiple rounds of yeast assembly. Clonal fragments are designed containing regions of homology for the assembly of intermediate clones containing inserts of ~20kb. The fragments designed for the extreme 5' and 3' end of the intermediate vectors are designed to contain embedded regions of homology, such that when intermediate inserts are excised, the final assembly can be constructed via a second round of yeast assembly in the final vector backbone. For yeast assembly we default to 70nt overlap between inserts to mediate the recombination events. Inclusion of this overlap can be automated in BOOST by increasing the overlap length in the partitioner tool. BOOST does not presently automate hierarchical assembly for sequential cloning, although this is planned as a future update. Sequence quality control for both yeast assembly and Gibson assembly constructs is performed on the Pacific Biosciences Sequel II platform.

Construction of libraries with high numbers of variants (typically thousands to hundreds of thousands of species) are ordered as mixed oligonucleotide pools. The complexity of individual pools are designed for hundreds to hundreds of thousands per pool depending on the requirements of the proposer's experiment. Since the oligonucleotides can be designed to be up to 300mers, there is sufficient allowance for 240nt of designed DNA plus 30nt either side to

mediate Gibson assembly into the destination vector. To generate double stranded DNA, yet minimize population skew, we subject each pool to 10 rounds of PCR amplification prior to cloning. Sequence quality control for highly complex libraries is performed on the Illumina Mi-Seq or Next-Seq platforms to determine library population and any possible bias.

Laboratory automation is fully embraced as part of project workflow, and comprises multiple liquid handlers and colony pickers, with all work being performed in either 96 or 384-well format. Our laboratory-automated apparatus presently comprises three liquid handlers and a colony picker. Specifically, we routinely utilize an 8-channel Biomek i5 liquid handler (Beckman Coulter), a QPix 460 colony picker (Molecular Devices) an Echo 525, and an Echo 550 (Labcyte, now Backman Coulter). All other apparatus comprises standard molecular biology laboratory apparatus, albeit compatible with 96- and 384-well format where appropriate (for example, PCR machines and centrifuges). Presently none of these devices are directly integrated with each other; instead a modularized approach is taken, with each machine being used as appropriate for any given project and plates manually transferred between machines.

Assembly for a given project is initiated only once the vector has been sequence validated and DNA building blocks and oligonucleotides have been delivered. We use Echo acoustic liquid handlers for the automated resuspension of oligonucleotides and DNA fragments as well as dispensing reagents for PCR mixtures. Successful amplification is manually confirmed by gel electrophoresis, reactions are cleaned in plate format and amplicons quantified by plate reader. Due to the low-throughput nature of agarose gels, typically only a subset of amplicons are analyzed by electrophoresis for large projects with more than a hundred inserts prior to proceeding to assembly. Gibson assembly reactions are scaled to 10µl final volume containing 30ng linearized vector. Reactions are established by Echo-automated transfer of reagents, into a 384-well PCR plate. After 1 hour at 50°C, 2µl of the reactions are transformed to in-house generated chemically competent cells, and cells plated to bioassay plates. Our default strain is

TOP10. 8 candidate colonies per insert are picked for overnight incubation, followed PCR-based validation of insert and preparation for sequence verification on the Pacific Biosciences Sequel II platform. For yeast assembly reactions, 100ng linearized vector and two times molar ratio of each insert fragment are combined with 25µl competent yeast cells, generated by standard yeast transformation methods, and plated onto selective media. Yeast colonies are screened by colony PCR, and positive colonies are lysed, and the lysate transformed to chemically competent *Escherichia coli.*

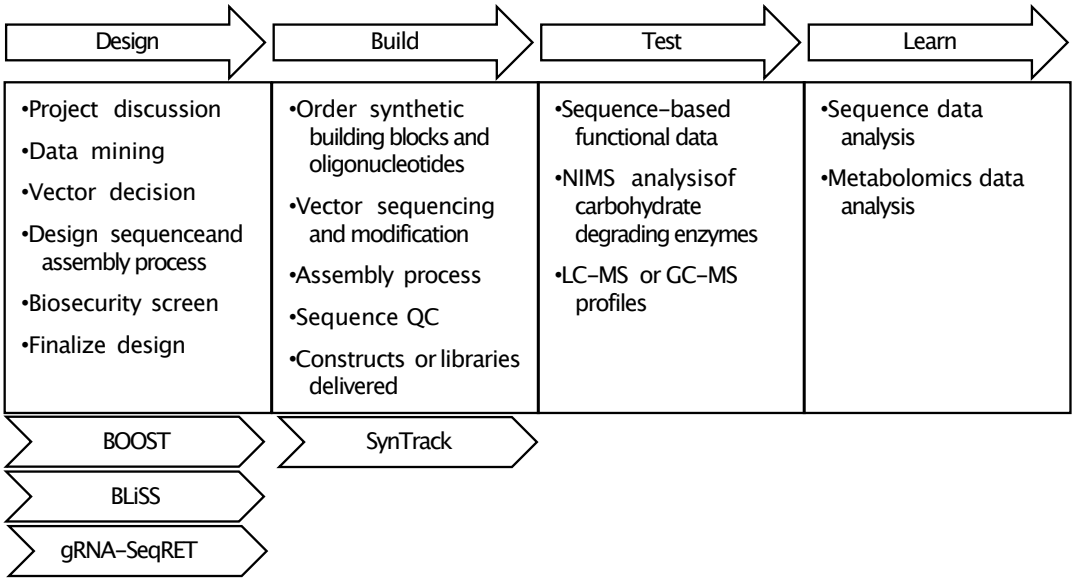
These general workflows are modified on a per batch basis to accommodate more complex assembly methods and other automation apparatus as needed. SynTrack tracks each step electronically as well as aiding progress by generating plate maps for resuspension of lyophilized reagents and final constructs. SynTrack additionally tracks the duration of each project, with the timeframe commencing with building block orders (since this occurs as soon as designs have been agreed upon and finalized) and ending with the construct shipment.

The platform aims to deliver at least 80% of requested constructs but averages >90% (note this varies with construct size and DNA sequence complexity. For example large constructs comprising multiple building blocks require 100% successful synthesis and delivery of constituent building blocks, and PCR-heavy projects are often outliers, bringing the mean down). The vendors we work with for providing synthetic DNA are selected on the basis of a Request For Proposals – in which companies compete on the basis of pricing, turnaround time, error rate and availability of APIs and eCommerce options. Not all requested strings are successfully synthesized by our primary vendor, in these cases, we have a secondary and (if necessary) a tertiary vendor.

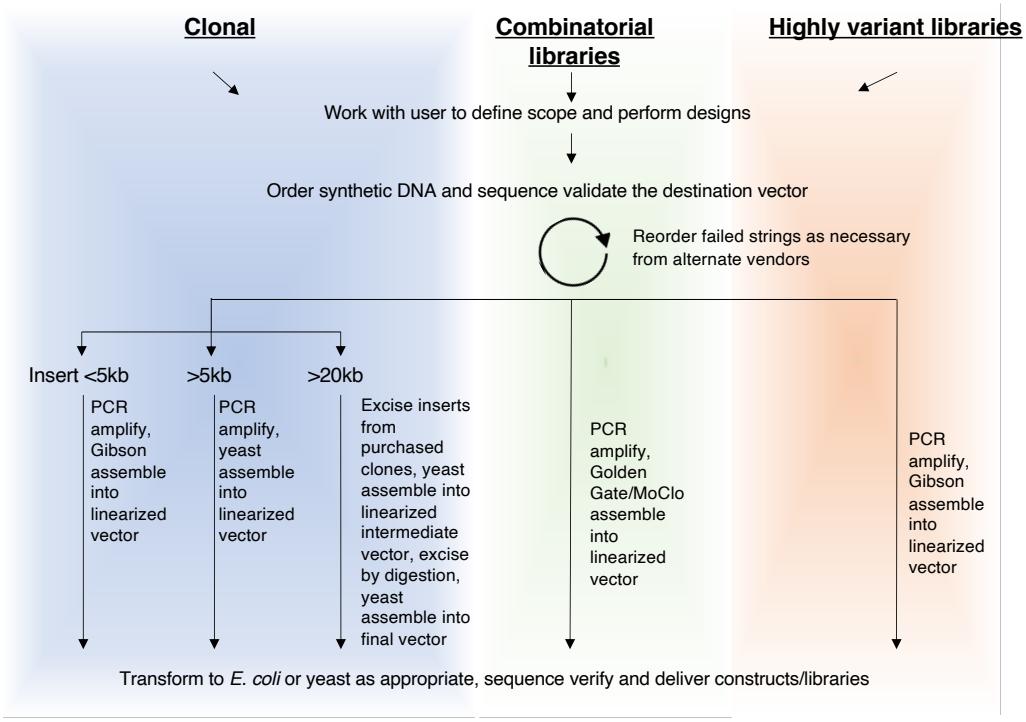
In our present workflow, some steps may be performed manually using multi-channel pipettes instead of automation for small numbers of assemblies or if a machine is in use. While affording

maximal flexibility with methods for assembly and general workflows, this approach both prevents maximal capacity from being achieved and necessitates constant human oversight. To help overcome this, one avenue that is presently being explored is a complete end-to-end integrated system for the automated assembly of constructs via Gibson assembly, for which the inputs would be DNA building blocks, oligonucleotides and reagents, and the output would be complete final constructs arrayed in plates.

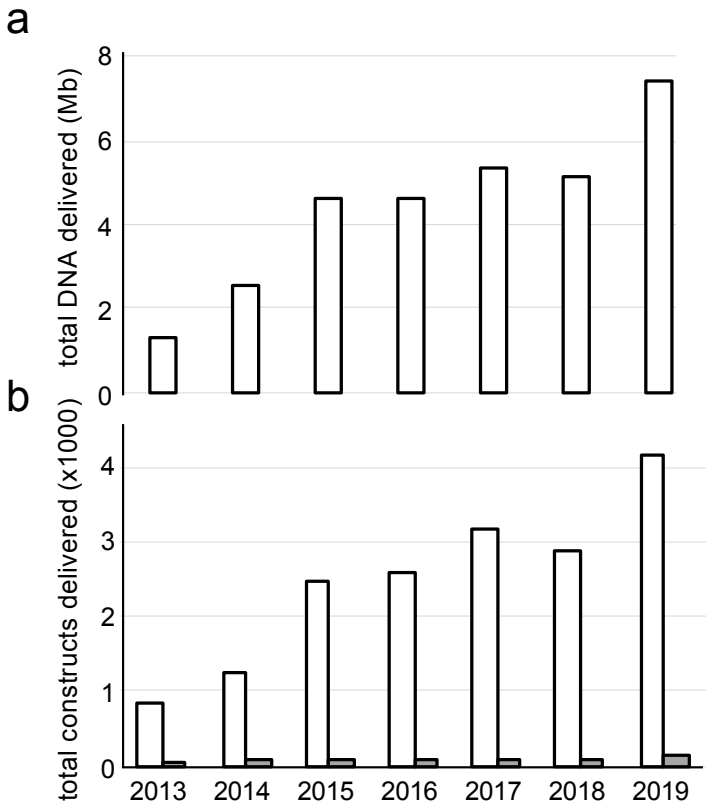

1. E. Oberortner, J. F. Cheng, N. J. Hillson, S. Deutsch, Streamlining the Design-to-Build Transition with Build-Optimization Software Tools. *ACS Synth Biol* **6**, 485-496 (2017).
2. U. Urquiza-Garcia, T. Zielinski, A. J. Millar, Better research by efficient sharing: evaluation of free management platforms for synthetic biology designs. *Synth Biol (Oxf)* **4**, ysz016 (2019).
3. N. Kouprina, V. Larionov, Transformation-associated recombination (TAR) cloning for genomics studies and synthetic biology. *Chromosoma* **125**, 621-632 (2016).




**Supplemental Figure 1. Overview of activities in the DNA synthesis and assembly platform at JGI.** The activities can be categorized into 4 stages that correspond to the Design–Build–Test–Learn phases of traditional engineering disciplines. All projects are initiated with a discussion between the project Principle Investigator (PI) and JGI staff to define the project scope and identify strategies to design and build constructs or libraries. In projects requiring data mining, domain experts of Phytozome, MycoCosm, IMG, and IMG/M participate in the discussion. Most user projects end after constructs or libraries are delivered at the end of the Build phase, with the user performing the work required for the Test and Learn phases. However, a limited number of projects continue with JGI input for the Test and Learn phases. This is usually for projects where the functional readouts are sequence or metabolite based. The informatic tools employed at each stage are shown beneath each phase, and include the design software BOOST and gRNA–SeqRET (design of constructs and libraries, respectively) BLiSS software (to flag sequences for biosecurity and export control concerns), and SynTrack (to track and provide workflow instructions for the laboratory processes.




**Supplemental Figure 2 Overview of molecular workflows.**




**Supplemental Figure 3 DNA synthesis and assembly.**

(a) As of 2019, approximately 7Mbp of DNA and assembled constructs are synthesized per annum. (b) Total number of small (<5kb) and large (>5kb) inserts assembled per year are shown in white and grey respectively